Tweets by @RickyPo

Richard Poynd
@RickyPo
New York Governor has vetoed S2890B that would have publishers "offer licenses for electronic books to libraries under reasonable terms." readersfirst.org/news/2022/1/3/…

NY G...
As firs...
reader...

6m

Richard Poynd
@RickyPo
Behind the scenes of our new open science website openworking.wordpress.com/2022/01/03/beh…

Behin...
We ha...

The State of Open Access
Predatory publishing
Institutional Repositories
Green OA **Gold OA** **Self-archiving**
Copyright
Basement Interviews
OA Interviews

Monday, June 02, 2014

Interview with Steve Pettifer, computer scientist and developer of Utopia Documents

Utopia Documents is a novel tool for interacting with the scientific literature. Developed in 2009, it is a free PDF reader that can connect the static content of scientific articles to the dynamic world of online content.

This week Utopia will be released as an open source project. It will also become the platform for a new crowdsourcing tool called Lazarus. With Lazarus, it is hoped to recover large swathes of the legacy data currently imprisoned in the charts, tables, diagrams and free-text of life science papers published in PDF files. This information will then be made available as an open access database.

The developer of Utopia is computer scientist Steve Pettifer, currently based at the University of Manchester. In a recent email conversation Pettifer explained to me the background to Utopia, and what he hopes to achieve with Lazarus.

Steve Pettifer

One of the long-standing debates within the open access movement is whether priority should be given to advocating for gratis OA (no cost access to read research papers), or libre OA (no cost access to read *plus* the right to reuse/repurpose papers).

Advocates for libre OA argue that since the benefits it provides are much greater than gratis OA, libre OA should be prioritised. Advocates for gratis OA respond that since gratis OA is achievable much more quickly and easily (and without additional cost to the research community), *it* should be prioritised. Besides, they add, very few researchers want to reuse research papers in any case.

In reply to this last point, libre OA advocates retort that the issue is not just one of reuse, but having the ability to text and data mine papers in order to create new services and databases and generate new knowledge. For this reason, they say, it is vital that papers are licensed under permissive Creative Commons licences that allow reuse (i.e. libre OA).

Passive reading

For similar reasons libre OA advocates dislike the widespread use of PDFs today. Designed to ensure that the (print-focused) layout of a document is the same whatever system it is displayed in, the Adobe Acrobat format is not conducive to text mining. So while it is fine for human readers, computers struggle to make sense of a PDF.

It may, for instance, not include information about who authored the document or the nature of the content in a form that machines can understand, since this would require the inclusion of metadata. While metadata can be inserted into PDF files, publishers/authors rarely go to the effort of inserting it. For this reason PDFs generally also do not have an explicit machine readable licence embedded in them to signal what can legally be done with the content.

In addition, any diagrams and charts in a PDF file will be static images, so machines cannot extract the underlying data in order to reuse or process the information.

- Home
- About Richard Poynder
- Blog: *Open and Shut?*
- The State of Open Access
- The Basement Interviews
- Open Access Interviews
- Essays on Open Access
- Archive

Blogs, News & More:
Interview 1: Richard Poynder
Interview 2: Richard Poynder
Interview 3: Richard Poynder
DOAJ
BASE
LENS
Digital Koans
LSE Impact Blog
Heather Morrison
The Scholarly Kitchen
Open Access India
PLoS Blogs
Redalyc
SPARC Europe
IP Watchdog

Search This Blog
Search

Popular Posts

Open Access: "Information wants to be free"?
(A print version of this eBook is available here) Earlier this year I was invited to discuss with Georgia Institute of Technology libraria...

PLOS CEO Alison Mudditt discusses new OA agreement with the University of California

OA Essays Open Access in Serbia Open Access in India Open Access in Egypt ScienceDirect Open Access in California OA in Latin America Open Access in the Humanities MDPI Preprints Selecting Reviewers Global Research Council OA Big Deal Open Notebook Science Elsevier Gates Foundation OA in South Africa OA in France SSRN OA & the Humanities Timothy Gowers Harold Varmus Peter Suber OA in Poland OA Embargoes Big Deal Finch Report Jeffrey Beall ALPSP OA Mandates PLOS Peer Review Springer BioMed Central Free Software Digital Preservation Dove Medical OA in Russia Radical OA Almost OA HEFCE Frontiers

Critics of the PDF also dislike the fact that it permits only passive reading. This means that scientists are not fully able to exploit the dynamic and linked nature of the Web. In fact, researchers often simply print PDF files out and read them offline. For these reasons, libre OA advocates, computer scientists, and forward-looking publishers (particularly OA publishers) are constantly trying to wean researchers off PDFs in favour of reading papers online in HTML.

Over a decade ago, for instance, the *Biochemical Journal* spent a great deal of time and effort revamping its site. It did this sufficiently well that it won the 2007 ALPSP/Charlesworth Award for Best Online Journal — on the grounds that it had successfully "overcome the limitations of print and exploited the flexibility of the digital environment".

But to the frustration of the journal's publisher — Portland Press — despite all its efforts scientists simply carried on downloading the papers as PDF files.

Researchers, it turns out, still much prefer PDFs.

The question is however: Do PDF files allow scientists to make best us of the Web? This thought occurred to Steve Pettifer in 2008, as he watched a room full of life scientists trying to combine the work of two separate labs by downloading PDFs, printing them off, and then rapidly scanning the information in them. Surely, he thought, this is not a very efficient way of doing science in the 21st Century?

Since Portland Press had reached the same conclusion it offered to fund Pettifer and his colleague Terri Attwood to come up with a solution that would combine the appeal, portability, and convenience of the PDF with the dynamic qualities of the Web.

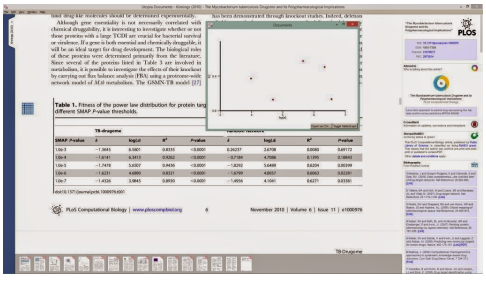
The outcome was Utopia Documents.

Utopia

Superficially, Utopia Documents is just another PDF reader. Unlike other readers, however, it comes with a number of novel interactive features. When a paper is loaded into it, for instance, a sidebar opens up on the right-hand side and fills with relevant data from external databases and services like Mendeley, SHERPA/RoMEO, and Wikipedia.

So, for instance, Utopia will pull the OA status of the paper and any self-archiving policy of the journal in which the paper was published from SHERPA/RoMEO. It also lists related papers from Mendeley; and where papers cited in the article are freely available on the Web it inserts live links to enable the user to download these papers into the viewer alongside the original paper. These can then be read/saved.

There is also a figure browser that lets the user flip through all the images in the document, and a page browser for jumping between pages.



Other interactive features include the ability to play and rotate molecules and protein sequences, and, where there is a reference to a drug, the molecular structure and formula of the compound can be pulled in.

In addition, it is possible to look up words and phrases in the paper by highlighting them and clicking "explore" from a popup menu. Amongst other things, this function allows structures from the protein databank to be pulled in, as well as associated laboratory products, related articles from PubMed, and news from SciBite.

The Public Library of Science (PLOS) and the University of California (UC) have today announced a two-year agreement designed to make...

P2P: The very core of the world to come

In the first part of this interview Michel Bauwens , the creator of The Foundation for P2P Alternatives , explained why he believes the var...

The OA Interviews: Taylor & Francis' Deborah Kahn discusses Dove Medical Press

Please note the postscript to this interview here The open-access publisher Dove Medical Press has a controversial past and I have writ...

The Open Access Interviews: Publisher MDPI

Headquartered in Basel, Switzerland, the Multidisciplinary Digital Publishing Institute, or more usually MDPI , is an open access publisher...

Community Action Publishing: Broadening the Pool

We are today seeing growing dissatisfaction with the pay-to-publish model for open access. As this requires authors (or their funders or ins...

Copyright: the immoveable barrier that open access advocates underestimated

In calling for research papers to be made freely available open access advocates promised that doing so would lead to a simpler, less cos...

The Open Access Interviews: OMICS Publishing Group's Srinu Babu Gedela

***Update: On August 26th 2016, the US government (Federal Trade Commission) announced that it has charged OMICS with making false claims, ...

Robin Osborne on the state of Open Access: Where are we, what still needs to be done?

In some cases it is also possible to manipulate and plot the data in the paper's tables as simple graphs, and to "play" 3D structures of proteins. Registered users can also comment on the paper in the hope of starting a conversation about it with other users.

So how does Utopia turn a static PDF file into a dynamic interactive document? Pettifer explains: "Where Utopia is able to find the machine-readable version of a paper online, either because it exists in a well-known open place (such as PubMed Central) or because a publisher has told us how to get hold of it, we can consume that version on the fly and use it to enrich the PDF version by finding references, extracting tables of data and so on."

He adds, "Where we can't find any semantic information inside the PDF itself (publishers rarely put this in) or online anywhere, Utopia tries to guess what's going on by reconstructing some of the semantics from the typographic layout of the article."

Pragmatic compromise

It is important to note, however, that Pettifer does not assume that Utopia is a long-term solution. Rather, he says, it is a pragmatic compromise in recognition of researchers continuing preference for PDFs.

Certainly we should not doubt that scientists continue to prefer the PDF. A survey undertaken by *Nature* earlier this year, for instance, suggested that the overwhelming majority of researchers still view the PDF as the best way to read scholarly papers. Specifically, 39.24% of those surveyed said they preferred to read academic papers in PDF on their desktop computer, 43.4% preferred printed PDF, and 11.28% preferred reading PDFs on a Tablet (giving a total of 93.92% who preferred the PDF format). Only 4.34% said they preferred HTML.

Nevertheless, suggests Pettifer, we can expect this to change — for a number of reasons. First, as more and more people start to read scientific articles on handheld devices the appeal of the PDF will likely wane. "A PDF can be quite pleasant to read on a big screen (centuries of typographic craft have gone into working out how to lay things out 'on paper'), but is often fairly awful on a tablet or mobile phone, especially if it's a two-column format."

Second, suggests Pettifer, the growth of OA will also likely change scientists' habits, since he suspects that the popularity of the PDF is partly a consequence of closed access publishing. "I'm fairly sure that part of the motivation for hoarding PDFs is that we know they can't then be taken away from us if we move institutions, or our subscriptions to a particular journal expire. I think that if I could be certain that an article was always accessible to me wherever I happened to be because it's open access, I'd personally be more comfortable with reading it on line rather than cluttering up my hard disc with files."

Third, Pettifer says, research papers will increasingly have to be read by machines rather than by humans; and, as we saw, PDF files are not exactly machine-friendly.

"We are already way past the point where any single person, or even research group, can hope to read all the relevant literature in even fairly niche areas of science, so we have to assume that the primary consumers of scientific 'papers' in the future are going to be machines," says Pettifer. "And that means that we have to create content that is suitable for them first and foremost."

Machines first

One of a series exploring the current state of Open Access (OA), the Q&A below is with Robin Osborne , Professor of Ancient History a...

The OA Interviews: Frances Pinter

In 2012 serial entrepreneur Frances Pinter founded a new company called Knowledge Unlatched (KU). The goal, she explained in 2013, was ...

Clearly, therefore, in the future papers will need to be comprehensible to machines as well as to humans. However, says Pettifer, this need not be problematic, so long as the needs of the machine are prioritised. "[M]achines can turn their formats into ones suitable for human consumption much more easily than going the other way round. So it really doesn't matter hugely whether we have PDFs and ePubs and Mobi and HTML and so on being created all at the same time, as long as these are all generated from the same machine-readable source."

One can therefore envisage a future in which the default format for research papers will be a machine-readable one, but it will be possible to create human-readable versions as and when needed — a variation on print-on-demand perhaps?

Given all this, we might be inclined to conclude that Utopia's usefulness will be short-lived. In reality, however, it may be that its real value has yet to be realised.

Consider, for instance, that the only people to read many research papers today will be the author, the editor and the reviewers. How many papers are never read by anyone other than this small group is a source of disagreement, but it is widely assumed that many papers are never downloaded and/or cited. As it happens, this may be the fate of a great many of the PDF files lying around on the Web. A recent report by the World Bank, for instance, concluded that nearly one-third of its PDF reports have never been downloaded. Another 40 percent have been downloaded fewer than 100 times, and only 13 percent have seen more than 250 downloads.

Commenting on the World Bank's finding on the *Washington Post* website Christopher Ingraham said, "Not every policy report is going to be a game-changer, of course. But the sheer numbers dictate that there are probably a lot of really, really good ideas out there that never see the light of day. This seems like an inefficient way for the policy community to do business, but what's the alternative?"

Loading the Web with PDF files of research papers that may never be read by other scientists would seem to be an equally inefficient way for the research community to do business.

But suppose there were an alternative? Suppose that all the data and good ideas in those PDFs could be extracted and put in in an OA database that scientists could search?

Lazarus

This, in fact, is the next step for Utopia. Thanks to a grant Pettifer and his colleagues received earlier this year from The Biotechnology and Biological Sciences Research Council (BBSRC) Utopia will become the platform for a new crowdsourcing project called Lazarus that aims to recover large swathes of legacy data buried in the charts, tables, diagrams and free-text in the multitude of life science papers that have been published. Once extracted, this information will be converted into processable data and placed in a searchable database.

The logic of doing this is surely sound. In its current form this information may not only never be read by researchers but, as things stand, it cannot be mined, validated, analysed or reused by machines either — unless, that is, someone was prepared to go to a great deal of effort to recover it manually.

So, for instance, small molecules in PDF files are typically represented as static illustrations, biochemical properties as tables or graphs, and any protein/DNA sequences are generally buried in the text. None of this is understandable to machines today. In addition, references and citations will likely be in arcane formats, and other objects of biological interest referred to by ambiguous names.

To release this data by hand would require re-typing figures from tables, checking citations in digital libraries, and redrawing molecules by hand etc. — a highly time-consuming task.

In theory, of course, mass-mining methods (text mining, optical recognition) could be used to automate the task. However, the technology for doing this is not yet sufficiently reliable to be used without human validation. In addition, the

licenses under which the articles are published will invariably prevent the "bulk processing" that this implies. Bottom line: without "human computation", this knowledge is destined to remain entombed in the literature for the foreseeable future.

Unless, that is, tools like Utopia are deployed. Currently, when a scientist loads a paper into Utopia the locked data is only freed temporarily, since the text mining and semantic analysis Utopia performs when someone reads a paper is thrown away when the article is closed. With Lazarus, however, all the non-copyrighted bits of data released when an article is read will be collected and put into a central repository by the scientists participating in the project.

Web-based observatory

And since this data will be unlocked on an individual-researcher basis, with users extracting only the facts, it cannot be said to infringe copyright. Explains Pettifer. "[W]e're not storing or redistributing any copyrightable material, just data and 'citations' in the form of nanopublications. And we're not 'bulk processing' anything at all; simply collecting together the insights of a lot of individual people and, with their permission, putting them in a place where others can find them."

Clearly, the secret will lie in persuading a sufficient number of researchers to take part. How popular has Utopia proved to date? Unfortunately, we don't know. "It's horribly hard to tell right now," Pettifer says, "Because it's a desktop application that doesn't phone home and there's no requirement to register the only thing we can go off is downloads, which have been steady at around 50k per year for the past couple of years."

By contrast, he adds, when Lazarus kicks off "we'll end up with much more meaningful statistics because user provenance will get recorded against behaviour."

It is worth stating that by taking part in the Lazarus project researchers will not only help themselves, but benefit their community too. To stress this point, and encourage scientists to take part, Pettifer plans to create a web-based observatory. This will also gather per-article metrics and observe and steer the crowd toward data-resurrection campaigns.

To further incentivise researchers to take part, Pettifer and his team are releasing the source code of Utopia Documents. This, says Pettifer, "opens up opportunities for collaboration with other open source projects, and could maybe even get the developer community interested enough to contribute new stuff."

If a searchable crowd-sourced resource like Lazarus had existed in 2008 then presumably the life scientists that Pettifer observed trying to combine the work of two labs could have achieved their objective both more quickly and more easily than printing out and scanning multiple PDF files.

It is also clear that libre OA has a great deal to offer the research community. On the other hand, of course, the way in which Utopia/Lazarus works suggests that gratis OA would be sufficient for Pettifer's purposes. Moreover, as a result of the Hargreaves rules (which came into force on June 1st) text mining will be less hampered by copyright restrictions in future — in the UK at least.

To find out more about Utopia, Lazarus, and Pettifer's views on the future of scholarly communication please read the interview below.

The interview begins

RP: Can you say something briefly about yourself and your research interests?

SP: I'm a computer scientist. My original research was in distributed virtual reality, trying to figure out how humans could inhabit a shared, consistent virtual environment when the network that connects them is relatively slow and unreliable compared with our experience of 'causality'. We developed the kinds of techniques that now make multi player console games commonplace. From there I moved into visualising scientific data, and started working with biophysicists and chemists to tackle some of the problems in those domains.

RP: How did your earlier research feed into your current research?

SP: What I realised working with life scientists is that they have a very different take on 'the literature' to computer scientists. Whereas we tend to use traditional publication as a reason to attend conferences and workshops and for putting achievements 'on record', in the life sciences the literature is a much more central part of the knowledge base, and it was quite an eye-opener to realise how crude the tools were for interacting with the literature.

So we started looking at whether any of the 'distributed visualisation' techniques we'd been working on from other areas could improve the way that scientists interact with scholarly articles. I guess the common theme here is trying to understand what humans are trying to achieve, and how the distributed nature of the internet can help or hinder this.

RP: Why Utopia Documents?

SP: Two things happened independently, but which jointly convinced me that a new tool was needed.

The first was that I was working with Prof Doug Kell, and taking part in a curiously named 'Yeast Jamboree'. Two independent labs had been developing a computational model of yeast glycolysis, and this was a get-together to merge the best bits to create a consensus model [1].

This was the first time I'd seen scientists interact with 'the literature' in anger; a room full of people, furiously downloading PDFs, scanning them for evidence that a certain bit of the model was good or bad, and cross-checking facts with other papers and online resources. There were PDFs, printed and electronic all over the place, and this convinced me that there must be a better way of doing things.

The second was that Prof Terri Attwood had been approached by Rhonda Oliver and Audrey McCulloch who at the time were working for Portland Press. They were interested in exploring whether anything could be done to improve the quality and utility of the auxiliary materials in their flagship publication, *The Biochemical Journal*.

Portland Press had put a lot of effort into building their own publishing platform (which I believe won the ALSPS/Charlesworth Award a while back), but in spite of all the technical innovations around the HTML view, they were still seeing users downloading static PDFs far more often than consuming the online content.

Rhonda and Audrey eventually took the brave move of funding a one year project to build a tool that would help bring PDFs to life by dynamically linking content to online tools and databases. And from that was born Utopia Documents.

RP: What year did development on Utopia start?

SP: 2008; we launched the software at an event in the British Library on December 10th 2009.

RP: Is there a company behind Utopia Documents and do you earn revenue from the reader, or is it exclusively a research project?

SP: Utopia was originally a research project funded by Portland Press, then Astra Zeneca and Pfizer. The IP is now owned by a spinout company called Lost Island Labs, and although it's legally a limited company we have mostly treated it as a research account into which we can put any money we make from selling bespoke versions of the software so that we can keep the team going 'between grants' (unlike publishing an 'idea' in the literature, software requires a certain amount of ongoing maintenance, and grant income is erratic, so we decided we needed some mechanism of smoothing the income out).

More recently we've been able to get a BBSRC grant for harvesting data from the literature that will sustain the team for another three years, so we're in reasonably good shape from that point of view. We also have a number of other plans for Utopia, some of which will help with funding.

For good or ill

RP: Ok, can you first say something about what Utopia Documents does?

SP: The premise of Utopia Documents is that for good or for ill, scientists still prefer PDFs to HTML representations of articles.

So rather than ignoring this fact and trying to force scientists to use a format they are still largely uncomfortable with, Utopia attempts to blend the best of 'the web' with some of the nice qualities of PDFs.

The main technical challenge is that although it's possible to encode much of the same 'semantic structure' in PDFs as it is in HTML, publishers generally don't do this. So to bring PDFs alive we had to develop algorithms that 'retro fit' much of the semantics by analysing the structure and content of a PDF article.

This involves a blend of algorithms from computer graphics and vision for recognising the 'typography' of an article, as well as text mining techniques for recognising it's content.

RP: So Utopia can link papers to external databases and to services like Mendeley, SHERPA/RoMEO, PubMed Central, SciBite and Wikipedia, all of which makes reading a paper a more dynamic experience. Additionally, where papers cited in the PDF file are available on an open-access basis, users can link directly to those articles. Apart from this last feature what, if any, additional features does open access enable Utopia to provide? To what extent (if any) was Utopia envisioned as an open-access tool?

SP: Apart from the ability to fetch papers automatically, Utopia's functionality relies less on open access and more on whether publishers have machine-readable versions of their articles in an accessible place (which could be behind a firewall, but accessible via IP authentication).

Where Utopia is able to find the machine-readable version of a paper online, either because it exists in a well-known open place (such as Pubmed Central) or because a publisher has told us how to get hold of it, we can consume that version on the fly and use it to enrich the PDF version by finding references, extracting tables of data and so on.

Other features, like browsing by image or being able to get data on biological terms that the user highlights are completely independent of publishers; all you have to have for those to work is the PDF itself.

Where we can't find any semantic information inside the PDF itself (publishes rarely put this in) or online anywhere, Utopia tries to guess what's going on by reconstructing some of the semantics from the typographic layout of the article (you can try out the PDF to XML conversion outside of Utopia here).

RP: Can you say something about the bespoke versions of Utopia and why they are needed? (Presumably these are for pharmaceutical companies like AstraZeneca and Pfizer?)

SP: There are two reasons that companies pay for the commercial version. The first is that the free version is quite promiscuous in terms of the relationships it has with other free-to-use online resources; so as you look up terms in papers, it's communicating with bio-databases such as UniProt or PDB to find definitions and information about biological entities in the paper you're reading.

For most users that's no problem — it's no different from interacting with those databases via the web — but for pharma companies, the mere fact that you're interested in a particular protein or drug is commercially sensitive.

So we create versions of the software that can work behind a company firewall, so that commercial users can be sure they are not leaving any fingerprints behind on the internet that could cause their companies problems with IP later on.

The other thing we do is create connections between Utopia and whatever in-house knowledge bases the companies have themselves; so when commercial users interact with articles they can see what their company knows about a particular drug/transporter/gene as well as information that's in the public domain.

Hamburger to cow

RP: I'm thinking that Utopia doesn't really solve the problem that is exercising many minds in the OA movement today — how to enable effective text and data mining?

SP: It certainly doesn't address the real problem, which is the licensing/legal one, but I think the software we've built for turning un-semantic PDFs into something a machine can get traction on goes a long way towards this.

It would be much better if the articles were just created with the right markup in the first place of course; turning a hamburger back into a cow is a rather messy business.

RP: The PDF format is increasingly frowned upon by scientists as a vessel for scientific papers (indeed, researchers have started to hold conferences with titles like "Beyond the PDF"). However, as you point out, for good or ill most still prefer to download PDFs. Perhaps it is for this reason that you describe Utopia as a compromise. But does this mean that you view Utopia as a stopgap solution until a better one emerges, or do expect to see more dynamic readers like Utopia being developed, and used on a long-term basis?

SP: I take issue with the idea that the PDF is frowned upon by scientists: there are definitely a very vocal few who really don't like it (and I completely understand why, and largely agree with them), but compared with the wider body of scientists it's a tiny minority.

If you do a show of hands at any scientific meeting outside of the Beyond The PDF / FORCE11 community, the vast majority — and I mean 90% or more — of scientists say they primarily interact with the literature by downloading PDFs; and even those that occasionally read or browse articles online will also snag a copy of the PDF version to keep for later reading and reference.

See, for example this recent poll here.

Most of the PDFs produced by today's publishers are really awful for machines to read (and that's as much because publishers don't use the features the format has to offer properly). As I indicated, that's a real impediment for people wanting to do data or text mining (though less of an impediment than the licensing issues that typically affect the same content). Which is why for Utopia to work we've had to go to the lengths of writing software to try to recreate the semantic structure of PDFs from their visual appearance.

We built Utopia as a pragmatic compromise; it would be great if we didn't have to deal with PDFs, but the reality is that for now we do.

The cracks are beginning to show for PDF as a human-readable format too, as more people shift to small form factor devices for reading scientific articles. A PDF can be quite pleasant to read on a big screen (centuries of typographic craft have gone into working out how to lay things out 'on paper'), but is often fairly awful on a tablet or mobile phone, especially if it's a two-column format.

So I think as these devices become more common, we will finally see a move away from the PDF towards something like ePUB or just HTML that can reflow sensibly to fit on a tiny screen.

RP: There is a good example of this transition here, where the main publishing format of the Code4Lib journal has been changed from PDF to EPUB. This was done, the publisher explains, because of the growing use of e-book readers plus the accessibility problems posed by PDF. In terms of other solutions, I suppose we are mainly talking about XML and RDF? But how would you distinguish what you are doing with Utopia from what eLife is doing with the Lens Project, Elsevier is doing with its Article of the Future and "executable papers" projects, and Wiley is doing with its Anywhere Article? Is it that these latter approaches (like that of Portland Press) focus on screen-based HTML solutions — a process I think is now called "enhanced publication", whereas Utopia uses a combination of extracting and importing metadata from machine-readable versions of papers elsewhere, combined with guesswork/reverse engineering?

SP: Yes; well mostly yes anyway! Publishers are definitely getting better at making the online HTML reading experience a pleasant one, but there's still a very long way to go. I'd say that *eLife* and *PeerJ* are way ahead of the game on this front; they have clean, responsive designs and have avoided cluttering up the article page with dross that I really don't care about when I'm doing 'proper'

deep reading. They produce excellent PDFs of their content too. eLife's lens is particularly cool; the Article of the Future on the other hand I find a very frustrating experience, especially on small screens, so even more of an incentive for me to grab the PDF.

In any case scientists are, somewhat ironically, often quite conservative in their adoption of new technology — there's still a lot of people who just print PDFs out on paper. So I think that although the PDF's life is limited, it'll be a quite a few years yet before it vanishes all together (and there's still the issue of dealing with all the pre-XML legacy content that exists only in PDF form).

What I have concluded recently though is that a big part of the popularity of PDF is a side-effect of closed access publishing. I've not done a proper study of this, but I'm fairly sure that part of the motivation for hoarding PDFs is that we know they can't then be taken away from us if we move institutions, or our subscriptions to a particular journal expire.

I think that if I could be certain that an article was always accessible to me wherever I happened to be because it's open access, I'd personally be more comfortable with reading it online rather than cluttering up my hard disc with files.

RP: Might it rather be that many remain far more comfortable reading research papers on paper? As you say, many scientists still print out PDFs and read them offline?

SP: I think there's some truth in that. I've largely weaned myself off printing articles to paper, but still find myself doing it very occasionally if I know I'm going to be doing a lot of annotation or want a really un-disturbed read.

But I'm pretty sure that eReaders and tablets will start to have an effect on this in a way that desktop reading hasn't so far.

RP: From what you say I assume Utopia is able to do more with papers from some publishers than with papers from others — where, for instance, the publisher encodes some semantic structure into its PDFs. Or is Utopia able to guess as accurately as if it had access to the metadata?

SP: It definitely varies; if we can find definitive metadata online then we can do more stuff more accurately. If we can't find it, then we have to apply heuristics to recover the semantics, and that can be hit and miss. Generally we're pretty good at it now, but we can't guarantee to get everything right all the time.

Largely automatic

RP: I understand the* Biochemical Journal *was marking up its papers for specific use in the Utopia Reader. Is it still doing this, and have other publishers begun to do so?

SP: Yes and no. Utopia's original emphasis was on editorial annotation of articles — something that boutique publishers like Portland can do because they have dedicated in-house editors, and that large-scale publishers such as Springer and Elsevier generally can't do because they rely more on academic good will.

Over the years as named-entity recognition and our semantic analysis of the article has improved we've found there's less of a benefit to the manual annotation in any case; so now it's largely automatic once we have found the article's metadata.

RP: Do you think the need for enhanced publication is more of an issue for scientific (STEM) researchers than it is for humanities and social science (HSS) researchers?

SP: I'm not sure I know the other fields well enough to comment really; the Life Sciences in particular are very well served by ontologies, databases and online tools, so Utopia works very well in that field.

Elsewhere there are fewer services we can call on, so less we can do with articles beyond extracting tables and references.

RP: You said Utopia was a "pragmatic compromise" and that you expect a move away from PDF going forward. Can you say more about how you see

scientific publishing developing in the future?

SP: My take on this is that as the amount of stuff getting published increases we are going to be more and more reliant on machines to help us figure out what to read; and that means they have to do some of the reading for us.

We are already way past the point where any single person, or even research group, can hope to read all the relevant literature in even fairly niche areas of science, so we have to assume that the primary consumers of scientific 'papers' in the future are going to be machines. And that means that we have to create content that is suitable for them first and foremost.

The nice thing about doing that (which for now means XML and RDF, but in the future could be some other machine-readable format-de-jour, it really doesn't matter too much) is that machines can turn their formats into ones suitable for human consumption much more easily than going the other way round.

So it really doesn't matter hugely whether we have PDFs and ePubs and Mobi and HTML and so on being created all at the same time, as long as these are all generated from the same machine-readable source.

RP: So you are saying that increasingly the focus of scholarly publishing is going to shift to creating machine-readable files from which PDFs and other human-friendly formats can be created on the fly (as and when needed) — which is the reverse of what Utopia currently does (cow to hamburger vs. hamburger to cow)?

SP: Essentially yes; at the moment Utopia has to guess at a lot of the semantics; if those were available in machine-readable form we wouldn't have to do that. Instead we could just 'project' those semantics onto whatever format the user happens to like, whether that's PDF, ePubs or whatever.

The other crucial aspect is not just shipping the text and images around in a machine-readable container, but of making the content amenable to machine processing; and that means as authors / editors / publishers we're going to have to get better at bridging the gap between the ambiguities of natural language (which is nice for humans to read) and formal notations that computers can make sense of.

At the moment these are worlds apart; computers struggle to make sense of scientific prose, and humans find it very hard to write things in a way a machine can interpret. I think that if we can't bring these things together somehow, we're going to be in trouble fairly soon.

The bigger problem

RP: Do you have a sense of how that might be done and what might hold it back?

SP: There's a really interesting bit of work going on in this space at the moment called the Resource Identification Initiative which is trying to encourage authors and publishers to include machine-readable identifiers in the text of their articles; it's very early days yet, but I think if it gets wider uptake it could be a really important step towards making the important bits of articles accessible to machines as well as humans.

There's nothing very technologically complicated involved here; the real problem as far as I can see is one of tradition and momentum. Part of the problem is that doing this kind of thing implies a cost to authors in return for a benefit to readers; it's easy for us to forget that every scientist wears both those hats at different times of the day.

The bigger problem is that today's publishing industry is based on a business model that relies on pretending that the internet doesn't exist. It made sense to charge for publishing as a service when you were laying things out with lead type and shipping them round the country on a horse and cart, but we all know that it's easy and cheap to get material out on the internet for people to read with very little technological expertise.

The only things the industry has left that are of any value are 'peer review' — and the value of that is subject of some debate — and the very questionable kudos associated with being published in a 'top journal'. I'm not saying that

‘self-publishing’ is the way forward, but the costs of ‘traditional publishing’ seem utterly out of kilter with the claimed added value.

I’m pretty sure at some point that funding bodies are going to call shenanigans on the cost of the current process, so it strikes me that if industry could position itself as doing something constructive that is a consequence of and in harmony with the existence of the digital media and the internet, rather than fighting against it, that would be a good thing.

RP: You are saying that instead of fighting the exigencies of the Internet publishers could be doing more to exploit its benefits. What sort of things do you have in mind?

SP: Helping authors make their material machine-readable. It doesn’t sound very sexy right now, but I think it will be crucial in the future, and not everyone wants to have to learn about identifiers and ontologies. I think that’s something where publishers could add real value to the material they process, rather than obsessing about typography and copyright.

RP: When I spoke recently with COAR’s Kathleen Shearer we discussed the discoverability problem arising from the failure to attach adequate metadata to papers posted in institutional repositories. We can expect this problem to be mitigated, she said, by the use of “new, automated methods for assigning metadata and repository software platforms can build-in standard vocabularies and metadata elements.” Earlier you said that much of what Utopia does now is automated. Would I be right in thinking that new methods are being (or will be) developed to automate the process of making research papers machine-readable in the first place? If so, this would make it much easier for authors, and presumably for publishers. But if that is right, what in your view would be the implications for publishers as they seek to retain a key role in the digital environment? The value they provide might seem to be becoming increasingly redundant.

SP: The metadata requirements for ‘discoverability’ and what I’ll loosely call ‘reproducibility’ or maybe ‘verifiability’ are quite different.

For discovering content you can afford to have things a bit off because the nature of the kind of query you’re going to be running is naturally a bit fuzzy (“find me all the papers to do with skin cancer”). For this, automated techniques can work pretty well, and people can cope with a few false positives and negatives much as they do with any online search.

But when it comes to specifying objects of interest, or being precise about claims that are made in a paper it’s much harder to develop automatic techniques able to guarantee accuracy. If a paper talks about a particular gene or antibody, it would be good to know that the link I have to a database entry is definitively to the thing the author meant, and not something that a heuristic has guessed might be right.

So computers are good at getting together a lot of data for broad overviews, but when it comes down to detail it’ll be quite some time before we can rely on them to be as accurate as a human. Of course to make the associations manually you need a human that understands the scientific domain, as well as how to represent this in terms of ontologies and identifiers and such, and who can confirm the result with the original author in order to get the provenance of how that link came to be made right! So it’s quite a specialised job.

I can imagine small or society-scale publishers maybe being able to do this; whether large throughput publishers could get together teams with the right experience at the right scale is another matter. But apart from this it’s hard to see what else publishers can provide that counts as real value in the long term (I’ll probably never get anything published again now).

RP: This is what I take from what you have said. Utopia does three things. 1. It turns static documents into dynamic documents. 2. It links the literature with the underlying data. 3. It provides a better way of meeting the needs of both machines and humans. 4. Something more drastic will need to be done in order to adapt the research paper to the Internet, but publishers are holding back the process of change. Is that correct?

SP: Yes that sums it up nicely.

RP: I see your new grant is for a project called Lazarus. This aims "to harness the crowd of scientists reading life-science articles to recover the swathes of legacy data buried in charts, tables, diagrams and free-text, to liberate process-able data into a shared resource that benefits the community." Is this connected with Utopia, or a development from it? Would you describe it as an open access project (since presumably the results will be freely available)?

SP: The plan is to use Utopia as a way of resurrecting the dead knowledge in the literature. At the moment all the text mining and semantic analysis that Utopia performs when someone reads an article gets thrown away when they close the article; the idea behind Lazarus is that non-copyrighted bits of data from the article will get pooled in a central repository that others can then search.

So it's a crowd-sourcing approach to generating a rich knowledge-base about the literature. All the data generated will be open access, and attributed both to the original source (so we expect it will drive traffic to the original article which will keep publishers happy) and to whoever contributed it.

Stories that persuade with data

RP: You said that the real problem with text mining is a licensing/legal one. We should perhaps highlight this, not least because it will presumably circumscribe what you are able to do with Lazarus.

You will also know that there is a long-standing debate within the open access movement between those who advocate for so-called gratis OA and those who insist that only libre OA will do. Essentially, gratis OA provides eyeball access to papers alone, whereas libre OA allows content to be repurposed and reused (by the use of Creative Commons licensing).

And the licensing issue is key since, as you pointed out, traditionally licensed and PDF-provided eyeball access to papers is fine for humans, but poses a serious problem for machines, and for people who want to text mine documents and reuse and repurpose scholarly papers as you plan to do with Lazarus. I assume, therefore, that you sit firmly in the libre OA camp?

SP: Very much so. I've not seen a single plausible argument (outside of 'protecting our business model') why anyone would want to publish a scientific article with any license other than CC-BY.

But there's an interesting point here; it's actually very hard to stop a single user from extracting knowledge from a single article at a time because it's hard to pin down what 'automated means' implies.

As soon as you open a PDF to read or print with any reader, you are processing it digitally (and to try to prohibit that would make the thing you've bought utterly unusable and therefore probably not a fair sale). I think it would be hard to object or prevent someone from using the 'find' function in a PDF reader to identify a particular term; and from there doing that repeatedly to find multiple terms isn't a huge step. It's certainly not illegal to then cite an article to say "I found the words *ibuprofen* and *inflammation* in this paper; therefore I think it's a paper about ibuprofen and how it affects inflammation" and maybe even to quote fragments of it to back your claim up.

And that's pretty much what we're doing in Lazarus; we're not storing or redistributing any copyrightable material, just data and 'citations' in the form of nanopublications. And we're not 'bulk processing' anything at all; simply collecting together the insights of a lot of individual people and, with their permission, putting them in a place where others can find them.

No scientist would trust my claim that this is a paper about ibuprofen and inflammation without wanting to read the full article in any case (maybe being directed automatically to the relevant parts via something like Utopia), so I think this will actually end up driving more traffic to full text articles to confirm interesting claims.

Moreover, my understanding is that the UK Hargreaves rules mean that the right to read is now the right to mine, in the UK at least. So the problem is beginning to ease.

RP: You said earlier that you had other plans for Utopia. This is in addition to Lazarus?

SP: Apart from the Lazarus activities, we've got a few new features in the pipeline that users have asked for and that we think will really transform the way Utopia works — but it's probably a bit too early to say too much about those right now.

The most exciting thing for us is that we're finally in a position to be able to release Utopia Documents as an open source project. As a primarily research-focussed group we've been wanting to do this for quite some time (from the beginning really), but didn't want to shoot ourselves in our collective feet in terms of financial sustainability; you can't keep on going back to funding bodies for more money to keep software alive, and we wanted to make sure that we had some way of keeping the research group together through funding droughts.

We've always had a model where the software has been free to use, even for commercial purposes, and we have been able to get enough income to sustain the work on the back of consultancy and bespoke versions for use inside commercial firewalls. Now that we've tried that model for a while it's pretty clear that making the source code open won't damage that source of funding and at the same time it opens up opportunities for collaboration with other open source projects, and could maybe even get the developer community interested enough to contribute new stuff.

RP: When I spoke to John Willbanks a few years ago he painted a picture of a future in which the Internet will also become a major platform for sharing the physical tools of science — e.g. cell lines, antibodies, plasmids etc. What struck me as particularly interesting here is that he suggested the ability to acquire these tools will be embedded directly into research papers. So if a reader of an open access paper wanted more detailed information on, say, a cell line, they would be able to click on a link and pull up information from a remote database. If they then decided they wanted to obtain that cell line from a biobank, they would order it in the same way as they might order an item on Amazon or eBay, utilising a 1-click system available directly from the article.

This suggests that the scholarly paper is set to become a great deal more than just a viewable document. Not only will it become the raw material for multiple machines and software agents to data mine, and the front-end to hundreds of databases (as we have discussed), but it will also be the launch pad for a range of ecommerce systems. The interview with Willbanks was six years ago and so thinking may have moved on. But what's your view on this?

SP: I think that's spot on. Anita de Waard described articles as 'stories that persuade with data', which I think encapsulates things nicely. Humans need the story and the ability to analyse the data; and as the scale of the literature increases we're going to be relying more and more on machines to help us do that. There's still a long way to go though.

RP: As I understand it, the research process consists primarily of gathering data, performing analyses and then sharing the results of those analyses with peers. This process has already begun to change as a result of the Internet and, as you say, it can be expected to change a great deal more in the future. But how might the process look in, say, 25 years' time? Will the research paper as we understand it have disappeared, and been replaced by something else altogether?

SP: The clearest description I've seen of this is what Prof. Dave de Roure and colleagues call 'Research Objects' — packages of knowledge that contain the story, as well as the data and any computational methods necessary to analyse or reproduce the claims being made. There's a lot of work going on at www.researchobject.org now that I think has really exciting potential. If anything is going to be the article of the future, I think it's research objects of some kind.

1. M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kürdar, M. Penttila, E. Klipp, B. Ø. Palsson, U. Sauer, S. G.

Oliver, P. Mendes, J. Nielsen, and D. B. Kell. *A consensus yeast metabolic network obtained from a community approach to systems biology*. *Nature Biotechnology*, 26:1155 - 1160, October 2008.

Posted by Richard Poynder at 09:12 

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)